一种基于梯度差的文档图像文本行检测算法

王 丹，王希常，杨 侠
(山东师范大学 信息科学与工程学院，山东 济南 250014)

摘 要：在分析文本行特点的基础上，提出了一种利用水平梯度差进行文档图像的文本行检测算法。该算法首先对输入的文档图像进行水平梯度差计算，然后在局部窗口中求解最大梯度差并进行文本行区域的合并，通过非文本区域过滤来消除字符阶跃的跳变，最后将文档图像以行块的形式进行显示。实验结果表明，与投影算法进行相比，该算法对于行间距较小的文档图像的检测效果较好，时间复杂度较低并且检测的正确率较高，具有一定的鲁棒性和较好的适应性。

关键词：梯度差；文本行检测；局部窗口；投影算法
中图分类号：TP751 文献标识码：A 文章编号：1674-7720(2011)18-0032-03

Text line detection of document images based on gradient difference

Wang Dan, Wang Xichang, Yang Xia
(School of Information Science and Engineering, Shandong Normal University, Jinan 250014, China)

Abstract: Based on analyzing characteristics of text lines, the algorithm of using horizontal gradient difference was presented to detect text line. Firstly, the horizontal gradient difference was calculated in the gray image, and then the maximum gradient difference was solved in the local window, after text line areas merging, document image shows as inline-blocks. In order to eliminate characters with different heights, the idea of non-text filtering was used in the paper. The result shows that this algorithm has high accuracy as well as good detection rate. In addition, the algorithm still has some robustness and better applicability.

Key words: gradient difference; text line detection; local window; projection algorithm

目前大多数文档图像的信息以数字化形式存在，并以文档形式存储在数据库中。文档图像处理是办公自动化的一个重要组成部分，在办公自动化、数字图书馆、图像视频检索等领域得到越来越广泛的应用[1]。其内容主要包括扫描输入、预处理、布局分析、字符识别等步骤，其中，文本行检测是进行布局分析、检索以及字符识别的重要组成部分。目前主要采用三种方法来进行文本定位：基于区域的方法、基于边缘的方法和基于纹理的方法[1]。基于区域的方法利用连通区域进行投影分析来获取文本区域，投影特性法[2]主要是对文档图像在指定方向上进行投影测试，根据投影的分布特征，在得到的结果中选取最佳的投影结果，以完成文本行的检测。但由于传统投影方法需要对整个图像进行指定方向上的投影，其计算量和复杂度都较高[3]。基于边缘的方法利用了图像中的文本与背景之间有较高对比度这一特性来进行定位。Chen Datong 等人[4]利用 Canny 算法提取图像边缘，并用形态学膨胀的方法将边缘连接成块，再利用基线定位完成文本行检测，但时间复杂度较高，当背景边缘较为复杂时，这类算法处理起来较为困难。基于纹理的方法利用文本具有的较强的纹理特征来区分背景，Mao Wenge 等人[5]利用小波变换检测图像纹理，再通过图像的纹理分析定位出文本。该方法通常具有较高的鲁棒性，但计算量大，复杂度较高，且文本定位不是很精确。

本文在总结上述算法特点的基础上，提出了一种基于梯度差的文本行检测算法，该算法利用了文档图像文本行特征，在水平方向上进行梯度差计算，然后进行文本行区域的合并和非文本区域的过滤，减少了文字粗细和图像分辨率的干扰，提高了检测的速度和精度。

1 文本行特点分析

文档图像文本行的特殊性主要表现在以下几个方面：

(1)大部分的文字边缘均突出，可以利用边缘信息进行文本检测，尤其是中文在水平和垂直方向上边缘均比较突出。边缘与梯度之间存在很大的关联，梯度的方向在数学中表示为某函数变化率最大的方向，在文档图像

中梯度往往反映了图像边缘清晰度[6],对于梯度较大的区域可表示为可能的文本区域。

(2)对于印刷体文档图像中的文本,同一行中文字的字符间距相同,间距与字符之间满足一定的比例关系,如字符间距大于字符宽度的1/5而小于字符宽度的两倍。在进行文本区扩展不同的字符区域使之成为一个有效的文本块时,非文本区域往往不具备该特征。对于手写体文档图像,字符间距不同,比印刷体文档图像复杂,但可以利用文本区域扩展特征进行文本行检测。

(3)文本行具有直线特征,有很强的方向性,可根据该特征进行文本行标记与定位,此外该特征还可用于倾斜校正和版面分析等。

文本梯度的信息不同于非文本区域的梯度,主要是由于一般文字和背景之间有很高的对比度。由于正负梯度值之差在文字区域较大,因此,本文利用梯度差方法进行文本行检测。

2 文本行检测算法

文本行检测算法没有进行文档图像的预处理过程,一定程度上减少了检测时间,如果输入的图像为真彩图像,首先进行灰度转化[7],这比单独对彩色图像的每个通道进行处理效率要高。

2.1 最大梯度差计算

字符图像往往具有较强的边缘信息,在字符边缘地带,相邻像素的灰度值变化剧烈,对应梯度幅度值较大。此外,文字行区域具有直线特点。因此,本文根据字符图像的特殊性,采用水平梯度差进行文本行区域的合并。其算法如下:

①对输入的文档图像 $I(x,y)$,利用滤波掩模[-1 0 1]进行卷积运算,得到梯度图像 G,计算公式如下:

$$|\nabla f(x,y)|=|f(x+1,y)-f(x-1,y)| \tag{1}$$

其中,$I(x,y)$为文档图像中的像素值。

②在一个大小为 $1\times w$ 的局部窗口内找出最大和最小梯度,二者的差值即为最大梯度差 MGD。计算公式如下:

$$G_{\min}(x,y)=\min_{x_i,y_i\in w(x,y)}(G(x_i,y_i)) \tag{2}$$

$$G_{\max}(x,y)=\max_{x_i,y_i\in w(x,y)}(G(x_i,y_i)) \tag{3}$$

$$MGD(x,y)=G_{\max}(x,y)-G_{\min}(x,y) \tag{4}$$

③根据梯度图像的像素平均值计算梯度图像的阈值 T:

$$T=\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}G(x,y)}{(m\times n-count)} \tag{5}$$

其中 $count$ 为梯度图像中大于平均梯度像素值的统计个数,$m\times n$ 为梯度图像 G 的大小。

④在局部窗口 w 中通过比较 $MGD(x,y)$和自适应阈值 T 的大小,得到二值化后的最大梯度差图像 $BMGD$,其中的每个像素值按照以下方法进行分类:

$$BMGD(x,y)=\begin{cases}1, & \text{if}(MGD(x,y)\geqslant T)\\ 0, & \text{if } (MGDB(x,y)<T)\end{cases} \tag{6}$$

2.2 文本行块标记

通常情况下,文档图像中的字符会存在字符高低不平的情况,为获取较为规则的文本行块,需进行消除字符阶跃的跳变。本文利用非文本过滤的基本思想,判断一个可能的文本区像素点两边是否满足非文本过滤的要求。主要方法是设定局部窗口,然后沿水平方向滑动,判断窗口内的像素是否全部为黑色像素(像素值为0),若满足,则停止计算,认为该区域为文本行区域,否则将窗口的像素值置为1。通过文本行定位可有效地消除字符间高低不平的情况,根据实际应用的需要,可再次进行非文本区域过滤操作,图1所示为输入的英文手写体文档图像,图2所示为文本行经过非文本区域过滤后得到的文本行检测效果。

图1 英文手写体文档图像

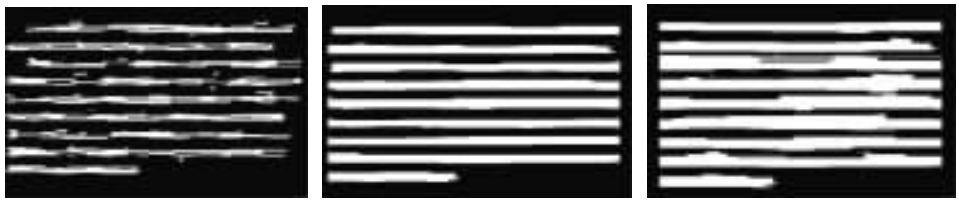
图2 文本行块检测结果

3 实验结果及分析

选择若干幅尺寸相同但字体不一的纯文字文档图像进行实验,实验环境为Windows XP操作系统,Pentium(R)1.7 G CPU,512 MB内存,用Matlab7.0仿真实现了文档图像的文本行检测。

经实验得到的阈值为77.5,为方便起见,本文选取80作为梯度图像的文本行检测阈值。在不同的局部窗口下对图1进行文本行检测,结果如图3所示。当局部窗口 w 取13时,行内会存在断点;当 w 取19时,看到行与行之间会有融合,二者效果都不理想;在 w 取得15时,效果较好。

将本文算法和投影检测算法[8]分别作用于印刷体文档图像中的某一图像(如图4所示),图5所示为利用水平梯度差得到的文本行检测效果,图6所示为利用投影算法得到的文本行检测效果。



(a) 窗口 $w=13$ (b) 窗口 $w=15$ (c) 窗口 $w=19$

图 3 不同局部窗口的文本行检测结果

图 4 英文印刷体文档图像

图 5 梯度差方法的文本行检测结果

图 6 投影检测算法得到的文本行检测结果

采用本文算法、投影检测算法分别对 10、20、30 幅图像分别进行实验，结果如表 1 所示。

表 1 本文算法与投影算法的平均检测时间对比

图像数	本文算法平均检测时间/s	投影算法平均检测时间/s
10	1.15	2.43
20	1.35	3.01
30	2.31	3.12

通过实验结果可以看出，在进行文本行检测时，对于行间距较小的文档图像，利用投影算法进行文本行检测时，行间距较小的文本行之间可能会发生融合，这样检测的正确率就会下降。本文算法通过最大梯度差和文本行标记算法可有效完成文本行的检测，且检测的平均时间短，因此具有较好的鲁棒性。

使用本算法对倾斜的文档图像(如图 7 所示)进行文本行检测，图 8 所示为文本行检测的结果。从图 8 可以看出，对倾斜的文档图像进行文本行检测时，会造成文本行融合现象，从而降低了检测正确率，这是本文算法的不足之处，需要进一步改进，以提高对倾斜文档图像的文本行检测正确率。

图 7 倾斜的文档图像

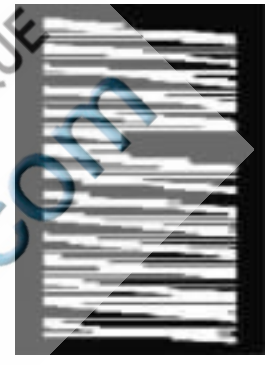


图 8 本算法的文本行检测结果

本文分析了文档图像的文本行特点，提出了一种基于梯度差的文档图像文本行检测算法，该算法计算简单、复杂度低。实验结果表明，该算法可以对印刷体以及手写体文档图像进行快速的文本行检测。本文算法也存在着不足，即在处理倾斜的文档图像时效果不佳，有待进一步改进。文本行检测算法可以为进一步进行文档图像的版面分析，深入进行文档图像检索、图文分割等奠定良好的基础。

参考文献

[1] 晋瑾，平西建，张涛. 图像中的文本定位技术研究综述[J].计算机应用研究，2007,24(6):8-11.

[2] 范玉凤.基于投影自适应算法的中文版面分析方法研究[J]. 光盘技术,2009(1):19-20.

[3] 吴涛,贺汉根.一种快速的文本倾斜检测方法[J].计算机工程与应用,2002:113-115.

[4] Chen Datong,SHEARER K,BOURLARD H. Text enhancement with asymmetric filter for video OCR[C].International Conference on Image Analysis and Processing,2001:192-197.

[5] Mao Wenge, Chung Fulai, LANM K,et al. Hybrid chinese/English text detection in images and vedio frames[C]. International Conference on Pattern Recognition,2002:1015-1018.

[6] 张弘.数字图像处理[M].北京:机械工业出版社,2007:115-118.

[7] JAE H K,TAE T P,YANG H C,et al. Photo-text segmentation in complex color document[C]. The 5th Japan-Korean Joint Symposium on Imaging Materials and Technologies, Kyoto, Japan,2004:44-47.

[8] Gao Feng,Zheng Nanning,Song Yonghong.Document images retrieval based on multiple features combination[C]. IEEE ICDAR, 2007.

(收稿日期:2011-06-17)

作者简介:

王丹,女,1987年生,硕士研究生,主要研究方向:数字图像处理,图像检索。

王希常,男,1967年生,研究员,主要研究方向:计算机图形学、数字图像处理。

杨侠,女,1985年生,硕士研究生,主要研究方向:数字图像处理。